Attribute	Guidance Bird
Needle bill shape	Needle bill shape
Spotted breast pattern	Brown Thrasher
Striped breast pattern	Song Sparrow
Solid tail pattern	Gray Catbird
Multi-colored tail pattern	Cedar Waxwing

Table 3. Guidance birds used the the generation of SUB.

A. Reference Birds

For SUB, we use the following 33 reference birds: Western Grebe, Black and white Warbler, European Goldfinch, Pacific Loon, White Pelican, Cedar Waxwing, Gadwall, Downy Woodpecker, Pileated Woodpecker, Purple Finch, Common Raven, White breasted Nuthatch, Northern Flicker, Mallard, Tropical Kingbird, Tree Swallow, Song Sparrow, Green Violetear, Gray Catbird, Green Jay, Cardinal, Red bellied Woodpecker, Pied Kingfisher, Rufous Hummingbird, Dark eyed Junco, Green Kingfisher, Horned Puffin, Anna Hummingbird, Barn Swallow, American Goldfinch, Lazuli Bunting, Blue Jay, Painted Bunting.

B. Guidance Birds

Guidance birds are used for pattern and shape modifications. We include in Table 3 the guidance birds chosen for each attribute when generating SUB.

C. Substitutions

We use the following list of substitutions in SUB: grey back color, grey bill color, white breast color, red breast color, blue breast color, grey crown color, white crown color, black crown color, pink crown color, yellow eye color, blue eye color, white eye color, grey forehead color, pink leg color, black leg color, grey leg color, green primary color, brown primary color, blue primary color, orange primary color, blue throat color, yellow throat color, green underparts color, red underparts color, white wing color, grey wing color, black wing color, spotted breast pattern, striped breast pattern, solid tail pattern, multi-colored tail pattern, and needle bill shape.

D. Prompts

A few example prompts:

$\mathcal{R}$ = European Goldfinch, $\mathcal{S}^+$ = Black crown color, $\mathcal{G}$ = bird, $c_{\mathcal{R}}$ = A photo of a European Goldfinch with black colored feathers on the crown of its head, $c_{\mathcal{G}}$ = A photo of a bird with black colored feathers on the crown of its head

$\mathcal{R}$ = Downy Woodpecker, $\mathcal{S}^+$ = Red breast color, $\mathcal{G}$ = bird, $c_{\mathcal{R}}$ = A photo of a Downy Woodpecker with a red colored breast, $c_{\mathcal{G}}$ = A photo of a bird with a red colored breast

$\mathcal{R}$ = Western Grebe, $\mathcal{S}^+$ = Solid tail pattern, $\mathcal{G}$ = Gray Catbird, $c_{\mathcal{R}}$ = A photo of a Western Grebe with a solid tail like a Gray Catbird, $c_{\mathcal{G}}$ = A photo of a Gray Catbird with a solid tail

$\mathcal{R}$ = Cardinal, $\mathcal{S}^+$ = Spotted breast pattern, $\mathcal{G}$ = Brown Thrasher, $c_{\mathcal{R}}$ = A photo of a Cardinal with a spotted belly like a Brown Thrasher, $c_{\mathcal{G}}$ = A photo of a Brown Thrasher with a spotted belly

$\mathcal{R}$ = Blue Jay, $\mathcal{S}^+$ = Needle bill shape, $\mathcal{G}$ = Hummingbird, $c_{\mathcal{R}}$ = A photo of a Blue Jay with the body of a Blue Jay and a beak like a Hummingbird, $c_{\mathcal{G}}$ = A photo of a Hummingbird

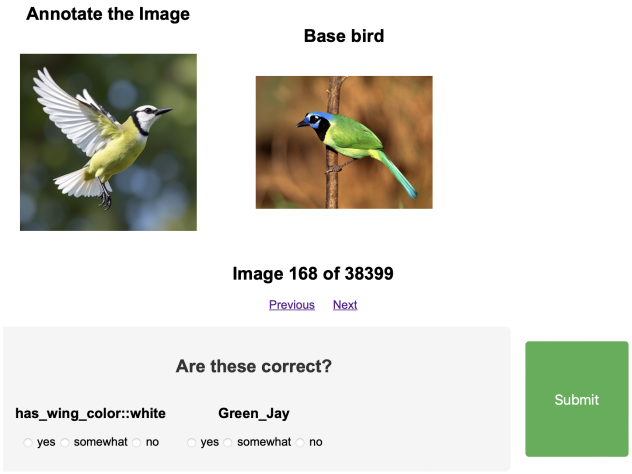

Figure 7. First human study interface with binary questions for attribute presence and reference bird faithfulness.

E. Human Verification User Interface

Human verification was completed by four volunteers. In Figure 7, we see the user interface used for our first user study, where participants were asked whether $\mathcal{S}^+$ was present and whether the bird accurately reflected the guide bird. In Figure 8, we show the interface for the second study, where the user is given all options in the target attribute group and asked to label which is present.

E.1. Reference Bird Verification

The underlying objective of specifying reference birds is to increase the overall diversity in birds exhibiting individual attributes. Specifically, we want to test the accuracy of attribute detection when it occurs in combinations not seen during test time. As long as $\mathcal{S}^+$ is present, it is not imperative that every synthetic bird closely match the reference class, but many should. As described in Section 4.3, we verify this on 40 images per attribute, by checking if the

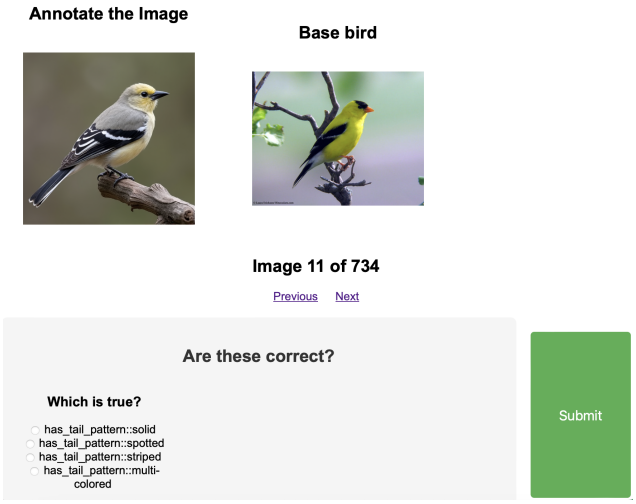

Figure 8. Second human study with attribute labeling within full attribute group.

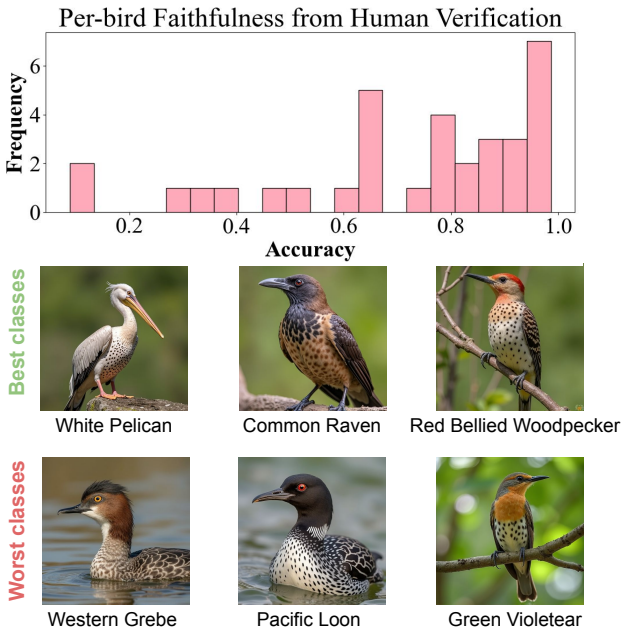

Figure 9. Histogram of the per-bird accuracy results from the human verification, where participants were asked if the attribute-substituted synthetic image represents the original bird. We see that 27 birds exceed 50% accuracy, showing that our generated dataset is very diverse.

synthetic bird is recognizable as the reference bird. We then calculate the percentage of faithful birds of all those generated for each bird class. In Figure 9, we show a histogram of this per-bird faithfulness. From this histogram, we can see that 27 out of 33 classes are faithful over half the time, and 10 classes are over 90% faithful. For the attribute *spotted breast*, we show examples from the three most faithful classes, and the three least faithful. While *Western Grebe*, *Pacific Loon*, and *Green Violetear* diverge from the representative class, we also note that they still provide some diversity to SUB.

F. VLM Random Chance Calculation and Label Set

For the VLMs, we calculate the probability of getting a single prediction correct at random if the target label is 1 as $\frac{1}{|\mathcal{A}|+1}$, where $\mathcal{A}$ is the attribute group corresponding to the target prediction and options $a_j \in \mathcal{A}$ are the manifestations of the attribute group. One is added to $|\mathcal{A}|$ to account for the additional option *none*. If the target label is 0, then it is $1 - \frac{1}{|\mathcal{A}|+1}$.

For SUB, we calculate the $\mathcal{S}^+$ random chance baseline across the modified attribute for each image in SUB, assuming a target label of 1. We calculate $\mathcal{S}^-$ from only the samples where the class-wise CBM label included a positive label for another attribute within the attribute group, and we consider that attribute with a target label of 0.

For CUB, we calculate the random chance baseline across all samples and CBM attributes, with the CBM class-wise labels as targets.

For selecting the possible label set $\mathcal{A}$ presented to the VLM, we use the full set of 312 CUB attributes for two reasons: (1) it offers a broader and more challenging set of plausible options than the CBM subset; and (2) the original labels used in CUB collection are well-aligned with expected dataset attributes, increasing the likelihood that the model selects the correct attribute over *none*.